Epidemiologic data analysis using R

Practicals 8

Faculty of Social Sciences, University of Tampere

–

Janne Pitkaniemi / Karri Seppä

23.3.2020

# Topics of practical 8

Learning objectives of this practical

* Estimating cumulative incidence function by using Aalen-Johansen estimator
* Testing differences in cumulative incidence by using competing risks regression model (Fine and Gray 1999) for subdistribution hazards

# Survival analysis: Oral cancer patients

## Description of the data

File oralca2.txt contains data from 338 patients having an oral squamous cell carcinoma diagnosed and treated in one tertiary level oncological clinic in Finland since 1985, followed-up for mortality until 31 December 2008. The dataset contains the following variables:

|  |  |
| --- | --- |
| variable | description |
| sex | a factor with categories: 1 = “Female”, 2 = “Male”, |
| age | in years at the date of diagnosing the cancer, |
| stage | TNM stage of the tumour (factor): 1 = “I”, …, 4 = “IV”, 5 = “unkn”, |
| time | follow-up time (in years) since diagnosis until death or censoring, |
| event | status at the end of follow-up (numeric): 0 = censoring alive, 1 = death from oral cancer, 2 = death from other causes. |

## Loading the packages and the data

Load the R packages Epi, mstate, and survival needed in this exercise.

library(Epi)  
library(mstate)

Loading required package: survival

library(survival)

Read the datafile oralca2.txt into an R data frame named orca.

Look at the head, structure and the summary of the data frame. Using function table() count the numbers of censorings as well as deaths from oral cancer and other causes, respectively, from the event variable.

orca <- read.csv("C:/Users/karri.seppa/Documents/TRE/oralca2.txt", sep="")  
head(orca)

sex age stage time event  
1 Male 65.42274 unkn 5.081 0  
2 Female 83.08783 III 0.419 1  
3 Male 52.59008 II 7.915 2  
4 Male 77.08630 I 2.480 2  
5 Male 80.33622 IV 2.500 1  
6 Female 82.58132 IV 0.167 2

## Lexis object with multi-state set-up

Before entering to analyses of cause-specific mortality it might be instructive to apply some Lexis tools to illustrate the competing-risks set-up.

Form a Lexis object from the data frame and print a summary of it. We shall name the main (and only) time axis in this object as stime.

orca.lex <- Lexis(exit = list(stime = time),   
 exit.status = factor(event,  
 labels = c("Alive", "Oral ca. death", "Other death")),  
 data = orca)

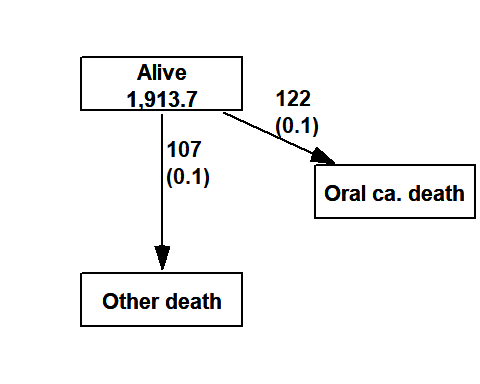
NOTE: entry.status has been set to "Alive" for all.  
NOTE: entry is assumed to be 0 on the stime timescale.

summary(orca.lex)

Transitions:  
 To  
From Alive Oral ca. death Other death Records: Events: Risk time:  
 Alive 109 122 107 338 229 1913.67  
   
Transitions:  
 To  
From Persons:  
 Alive 338

Draw a box diagram of the two-state set-up of competing transitions. Run first the following command line

boxes( orca.lex,boxpos=T )



Now, move the cursor to the point in the graphics window, at which you wish to put the box for “Alive”, and click. Next, move the cursor to the point at which you wish to have the box for “Oral ca. death”, and click. Finally, do the same with the box for “Other death”. If you are not happy with the outcome, run the command line again and repeat the necessary mouse moves and clicks.

## Event-specific cumulative mortality curves

We move on to analysing cumulative mortalities for the two causes of death separately, first overall and then by prognostic factors.

Use function Cuminc() in package mstate and view the structure of the thus created object.

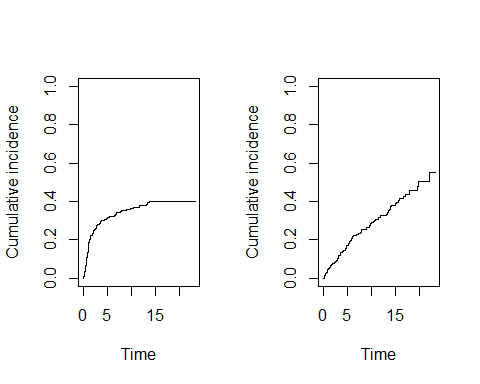
cif1 <- Cuminc( time = "time", status= "event", data = orca)  
str(cif1)

Classes 'Cuminc' and 'data.frame': 160 obs. of 7 variables:  
 $ time : num 0.085 0.162 0.167 0.17 0.246 0.249 0.252 0.329 0.334 0.413 ...  
 $ Surv : num 0.994 0.988 0.976 0.97 0.967 ...  
 $ CI.1 : num 0.00592 0.01183 0.01775 0.02071 0.02367 ...  
 $ CI.2 : num 0 0 0.00592 0.00888 0.00888 ...  
 $ seSurv: num 0.00417 0.00588 0.00827 0.00922 0.00965 ...  
 $ seCI.1: num 0.00417 0.00588 0.00718 0.00775 0.00827 ...  
 $ seCI.2: num 0 0 0.00417 0.0051 0.0051 ...  
 - attr(\*, "survfit")=List of 21  
 ..$ n : int 338  
 ..$ time : num 0.085 0.162 0.167 0.17 0.246 0.249 0.252 0.329 0.334 0.413 ...  
 ..$ n.risk : int [1:251, 1:3] 338 336 334 330 328 327 326 323 322 321 ...  
 ..$ n.event : int [1:251, 1:3] 0 0 0 0 0 0 0 0 0 0 ...  
 ..$ n.censor : int 0 0 0 0 0 0 0 0 0 0 ...  
 ..$ pstate : num [1:251, 1:3] 0.994 0.988 0.976 0.97 0.967 ...  
 ..$ p0 : num [1:3(1d)] 1 0 0  
 .. ..- attr(\*, "dimnames")=List of 1  
 .. .. ..$ : chr "(s0)" "1" "2"  
 ..$ cumhaz : num [1:251, 1:2] 0.00592 0.01187 0.01786 0.02089 0.02394 ...  
 .. ..- attr(\*, "dimnames")=List of 2  
 .. .. ..$ : NULL  
 .. .. ..$ : chr "1.2" "1.3"  
 ..$ std.err : num [1:251, 1:3] 0.00417 0.00588 0.00827 0.00922 0.00965 ...  
 ..$ sp0 : num 0 0 0  
 ..$ logse : logi FALSE  
 ..$ transitions: 'table' int [1:3, 1:3] 122 0 0 107 0 0 109 0 0  
 .. ..- attr(\*, "dimnames")=List of 2  
 .. .. ..$ from: chr "(s0)" "1" "2"  
 .. .. ..$ to : chr "1" "2" "(censored)"  
 ..$ conf.int : num 0.95  
 ..$ conf.type : chr "log"  
 ..$ lower : num [1:251, 1:3] 0.986 0.977 0.96 0.953 0.949 ...  
 ..$ upper : num [1:251, 1:3] 1 1 0.993 0.989 0.987 ...  
 ..$ conf.type : chr "log"  
 ..$ conf.int : num 0.95  
 ..$ states : chr "(s0)" "1" "2"  
 ..$ type : chr "mright"  
 ..$ call : language survfit(formula = Surv(time, statuscr) ~ 1, data = tmp)  
 ..- attr(\*, "class")= chr "survfitms" "survfit"

Function Cuminc() thus creates an ordinary data frame with quite self-explanatory column names. Unfortunately, no handy plot method is provided in the package, but in Epi package there is funciont plotCIF

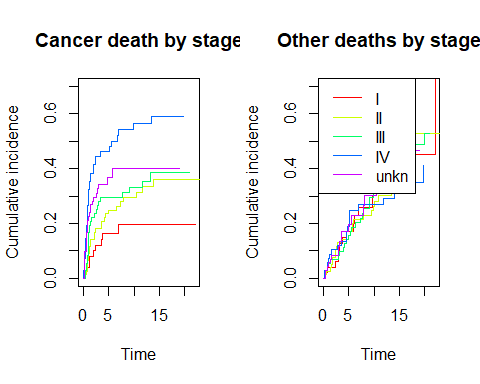
Draw two parallel plots describing the overall cumulative incidence curves for both causes of death

par(mfrow=c(1,2))  
cif1 <- survfit( Surv( time, event, type="mstate") ~ 1,  
 data = orca.lex)  
plotCIF(cif1,event=1)  
plotCIF(cif1,event=2)



Compute the estimated cumulative incidences by stage for both causes of death. Now you have to add argument stage when calling Cuminc(). See the structure of the resulting object, in which you should observe the first column containing the grouping variable. Plot the pertinent curves in two parallel graphs. Cut the -axis for a more efficient graphical presentation

par(mfrow=c(1,2))  
cif2 <- survfit( Surv( time, event, type="mstate") ~ stage,  
 data = orca.lex)  
 plotCIF(cif2, 1, main = "Cancer death by stage",  
 col=rainbow(5), ylim = c(0, 0.7) )  
 plotCIF(cif2, 2, main= "Other deaths by stage",  
 col=rainbow(5), ylim = c(0, 0.7) )  
legend("topleft",levels(orca.lex$stage),col=rainbow(5),lty=1)



Compare the two plots. What would you conclude about the effect of stage on the two causes of death?

## Competing risks regression

Test if the cumulative incidence differs between stages I and III by using the logrank test based on subdistribution hazards.

library(cmprsk)  
orca2 <- subset(orca, stage%in%c("I","III"))  
orca2$stage <- factor(orca2$stage)  
with(orca2, cuminc(time,event,group=stage)$Tests)

stat pv df  
1 3.6257709 0.05689121 1  
2 0.0200722 0.88733563 1

Test the difference based on the Fine-Gray model where age at diagnosis is or is not included.

Function crr fits the ‘proportional subdistribution hazards’ regression model. It It does not support the use of model formulas. Therefore, we fit a Cox model and use the corresponding design matrix in crr function.

orca2$agegr <- cut(orca2$age,c(0,55,75,Inf),right=F)  
  
# fit Cox model in order to   
fit.cox <- coxph(Surv(time,event%in%c(1))~stage,data=orca2)  
fit.finegray <- with(orca2,crr(ftime=time, fstatus=event, cov1=model.matrix(fit.cox),failcode=1,cencode=0))  
summary(fit.finegray)

Competing Risks Regression  
  
Call:  
crr(ftime = time, fstatus = event, cov1 = model.matrix(fit.cox),   
 failcode = 1, cencode = 0)  
  
 coef exp(coef) se(coef) z p-value  
stageIII 0.728 2.07 0.383 1.9 0.057  
  
 exp(coef) exp(-coef) 2.5% 97.5%  
stageIII 2.07 0.483 0.978 4.38  
  
Num. cases = 122  
Pseudo Log-likelihood = -153   
Pseudo likelihood ratio test = 3.86 on 1 df,

fit.cox <- coxph(Surv(time,event%in%c(1))~agegr+stage,data=orca2)  
fit.finegray <- with(orca2,crr(ftime=time, fstatus=event, cov1=model.matrix(fit.cox),failcode=1,cencode=0))  
summary(fit.finegray)

Competing Risks Regression  
  
Call:  
crr(ftime = time, fstatus = event, cov1 = model.matrix(fit.cox),   
 failcode = 1, cencode = 0)  
  
 coef exp(coef) se(coef) z p-value  
agegr[55,75) -0.157 0.854 0.428 -0.368 0.710  
agegr[75,Inf) 0.886 2.424 0.470 1.884 0.060  
stageIII 0.888 2.430 0.407 2.182 0.029  
  
 exp(coef) exp(-coef) 2.5% 97.5%  
agegr[55,75) 0.854 1.170 0.369 1.98  
agegr[75,Inf) 2.424 0.412 0.965 6.09  
stageIII 2.430 0.412 1.095 5.39  
  
Num. cases = 122  
Pseudo Log-likelihood = -150   
Pseudo likelihood ratio test = 10.1 on 3 df,

After adjusting subdistribution hazard for age the difference in cumulative cancer mortality between stages I and III is significant.